

A Unified Community Detection, Visualization and Analysis method

Michel Crampes* and Michel Plantié†

Ecole des Mines d'Ales, Parc Georges Besse, 30035 Nîmes Cedex

Abstract

Community detection in social graphs has been attracting researchers' attention for a long time. With social networks now so widespread on the Internet, community detection has recently become a major field of research. Most contributions focus on defining algorithms that optimize the so-called "modularity function". Interest was initially limited to unipartite graph inputs and partitioned community outputs. More recently, bipartite graphs, directed graphs and overlapping communities have all been investigated. Few contributions however have encompassed all three types of graphs simultaneously. In this paper, we present a method that unifies community detection for these three types of graphs while at the same time merges partitioned and overlapping communities. Moreover, the results are visualized in a way that allows for analysis and semantic interpretation. For validation purposes, this method is experimented on some well-known simple benchmarks and then applied to real data in three cases. In two examples of photo sets with tagged people, social networks are revealed. A second type of application is of particular interest herein. After applying our method to Human Brain Tractography Data provided by a team of neurologists, clusters of white fibers are produced in accordance with other well-known clustering techniques. Our approach to visualizing overlapping clusters also provides a better understanding of the neurological team's results, which lead to the possibility of applying community detection methods to other fields, such as data analysis with original enhanced performances.

*Electronic address: michel.crampes@mines-ales.fr; URL: <http://www.lgi2p.ema.fr/~crampes>;

†Electronic address: michel.plantie@mines-ales.fr; URL: <http://www.lgi2p.ema.fr/~plantie>

I. INTRODUCTION

Social group analysis is a long-standing field of research, particularly in the social sciences. [30] reveals that some studies date back as early as the 19th century; moreover, this review identified 21 results relative to the unique case of the Southern Women Dataset. Thanks to the Internet and growth in online social networks, community detection has become a major field of research in computer sciences. Many algorithms have been proposed (see several surveys on this topic: [12, 33, 34, 41]). These algorithms may be classified into four dimensions:

- The first dimension is dedicated to input data types, which may be:
 - unipartite graphs (nodes belonging to a single class),
 - bipartite graphs (nodes belonging to two classes, with no edges between nodes of the same class),
 - multipartite graphs (more than two classes, with no edges between nodes of the same class),
 - directed graphs (whereby the edges are oriented) - moreover, the edges may be weighted or unweighted.
- The second dimension is important since it pertains to the capacity of algorithms to comply with the experimenter's objectives, which may consist of a predefined number of communities (vs. a preference to determine this number after performing the math), a final partitioning (with every element belonging to a unique community), or overlapping communities (with some elements potentially belonging to several communities).
- In the third dimension, the algorithms are classified by performance, e.g. execution speed, data volume, stability, determinism vs. heuristics.
- The fourth and final dimension is scarcely mentioned in the various contributions. Since group detection is called upon by other scientists or for purposes of industrial data analysis, it is of utmost importance that the results be validated by the 'customer' and eventually analyzed and interpreted. This last objective requires both the algorithm and the entire methodology, from inputs to results presentation, to be easy to understand and implement from a variety of perspectives.

For this paper, the last requirement cited above will serve as a guide for proposing a unifying method applicable to community detection, visualization and analysis. Our first contribution will unify the three types of graphs listed above. This step will entail reducing bipartite graphs and directed graphs into unipartite graphs. The notion of building bridges from one context to another (e.g. from directed graphs to bipartite graphs) has already been explored by other authors. To the best of our best knowledge however, these attempts have only been piecemeal and, moreover, have not yielded any visualization or analysis tools. With this specific goal in mind, we will introduce a simple model of unified modularity for bipartite and directed graphs that differs from other authors' models. Ours will be formally derived from the typical GN modularity model for unipartite graphs [14], whereas other authors have used GN modularity for the a priori building of a specific model for bipartite graphs. These differences will be discussed below. We will then be able to extract partitioned communities from unipartite, bipartite or directed graphs with any algorithm that accepts unipartite graphs as inputs. In this paper, we will apply the Louvain algorithm [4], which is known for its efficiency in producing partitioned communities from extensive datasets. This algorithm is also applicable to weighted and unweighted graphs.

Our second contribution relates to the merging of partitioned and overlapping communities. Most authors propose models that extract either partitioned or overlapping communities, with the former being more frequent. According to our approach, when inputs are bipartite graphs, both node classes share the same resulting communities. Though the fact of having the same number of communities for both classes has been criticized by some authors [39], we will show herein that such a condition is relevant and especially useful for validating and analyzing the resulting communities. It is indeed possible to define overlapping communities in both classes using different belonging functions. From this result, we will present a model that merges partitioned and quantified overlapping communities, in a way that offers a unique view, as supported by a matrix visualization method that is both easy to produce and capable of facilitating analysis and interpretation.

In a third contribution, we will demonstrate how this composite view is of special interest in analyzing the detected communities. Other authors typically analyze their results through a comparison with other authors' results. While this is an option with our approach, it is also possible to observe different features of the extracted communities, then analyze them and create the potential for knowledge extraction.

For validation and comparison with other authors, our entire method has been experimented on small traditional unipartite and bipartite benchmarks. We have generated some interesting insight, which extends beyond previously reported results. For instance, when applying the Louvain algorithm [4] to detect communities in either bipartite or unipartite graphs, it is possible to observe local modularity optima and propose hypotheses that explain their origins. We can then apply our method to real medium-sized bipartite graphs, in a step that reveals significant properties such as community compactness and the role of inter-community objects. These results are valuable when observed in data like the people-photo datasets targeted by our experiments.

Beyond community detection, our method has been applied to brain data extracted through 'tractography' by a team of neurologists and psychoneurologists seeking to extract macro connections between different brain areas. Our results were compared with other results they obtained when applying traditional data analysis methods, and our community detection analysis tools led them to question their datasets and consider new hypotheses on brain connections. These two experiments are described in the paper in noting the following consideration. When working with people who are not specialists in the field (e.g. neurologists), it is of utmost importance that we justify our results and, above all else, explain any unexpected observations. It may also be necessary to improve results when 'greedy algorithms', such as Louvain's, might classify some people too quickly or more generally place objects into communities. We will show in this paper that our method is especially well suited to address such issues and may contribute to a better detection and understanding of communities or clusters in complex networks.

The next section will present a state-of-the-art on community detection techniques using different types of graphs. Section III will follow by focusing on a new method to unify all types of graphs. Section IV will then display how our unifying method is particularly valuable in computing and analyzing overlapping communities. In section V will discuss the visualization issue as applied to overlapping communities, while section VI will provide several practical results on various types of graph datasets.

II. BIPARTITE GRAPHS AND SOCIAL NETWORK DETECTION, STATE OF THE ART

A. Unipartite graph partitioning.

As stated above, several state-of-the-art assessments have already addressed the community detection problem [12, 33, 34, 41]. These have mainly focused on unipartite graph partitioning, i.e. each individual belonging to just one community. The calculation performed is based on maximizing a mathematical criterion, in most cases modularity [29], representing the maximum number of connections within each community and the minimum number of links with external communities. Various methods have been developed to identify the optimum, e.g. greedy algorithms [27, 31], spectral analysis [28], or a search for the most centric edges [29]. One of the most efficient greedy algorithms for extracting partitioned communities from large (and possibly weighted) graphs is Louvain [4]. In a very comprehensive state-of-the-art report [12] other newer partitioned community detection methods are described.

B. Extracting overlapping communities from unipartite graphs.

The partitioning of communities, despite being mathematically attractive, is not satisfactory to describe reality. Each individual has 'several lives' and usually belongs to several communities based on family, professional and other activities. For example, researchers may be affiliated with several partially overlapping communities when considering their scientific discipline, priority interests and scientific monitoring activities. As such, a greater number of methods take into account the possibility for overlapping communities. The so-called "*k-clique percolation method*" [32] detects overlapping communities by allowing nodes to belong to multiple k -cliques. A more recent method adapted to bipartite networks, based on an extension of the k -clique community detection algorithm, is presented in [38]. Several methods find overlapping communities through local fitness optimization [20][19]. The "Label Propagation Algorithms" (LPA) are efficient methods for detecting overlapping communities, e.g. [15]. [20] uses a greedy clique expansion method to determine overlapping communities, via a two-step process: identify separated cliques and expand them for overlapping by means of optimizing a local fitness criterion. Some research has provided results in

the form of hypergraph communities, such as in [7, 8]. Other methods are found in scientific papers, yet most of these are prone to major problems due to high computational complexity. More recently, Wu [40] proposed a fast overlapping community detection method for large, real-world unipartite networks. [9] derives n order clique graphs from unipartite graphs to produce partitioned and overlapping communities using Louvain algorithm. Interestingly with this method the number of communities is parameterized by the order n .

C. From unipartite to bipartite graphs.

The vast majority of community detection techniques are not really inspired by the meaning (semantics) of the relationship between nodes. When considering "semantics", it becomes necessary to focus on bipartite or "multi-partite" graphs, i.e. graphs whose nodes are divided into several separate subsets and whose edges only link nodes from different subsets. One example of this type of graph is the set of photos from a Facebook account along with their 'tags' [23] or else the tripartite network of epistemic graphs [36] linking researchers, their publications and keywords in these publications. Mining communities are often formed by converting a multipartite graph into a monopartite graph, through assigning a link between two nodes should they share a common property. Guimera [17] offered a modularity measurement for bipartite graphs, even those using a weighting parameter based on the number of shared properties, and then reduced the problem to a classical graph partitioning. In doing so however, semantics are lost; hence, many researchers retain the multiparty graph properties by extending the notion of modularity to these types of graphs [29] or else adapting the algorithms originally designed for unipartite graphs [2, 10, 24, 26, 39]. In [22] Liu and Murata presented a new and efficient algorithm based on LPA for bipartite networks.

D. Analysis and interpretation tools.

An important feature of the bipartite graph is its semantic value. A number of authors have developed methods to analyze and interpret the community results. Suzuki [39] compared modularity measurements stemming from various detection methods on well-known examples. This analysis proved limited however because it was exclusively quantitative and

did not take semantics into account. Moreover, the analyses in [22] were limited to modularity optimization, wherein detected communities from several authors were compared on the basis of their inner stability on standard benchmarks. Wu [40] also focused on a quantitative evaluation in terms of both modularity measurement and computation time. In all these contributions, the quality analysis has been limited to a modularity assessment, while a semantic analysis has hardly ever been considered. In contrast, we will show in the following section that our method provides for a deep interpretation of community membership, even though quantitative considerations (namely computational performance and modularity optimization) are still being included.

III. UNIFYING BIPARTITE, DIRECTED AND UNIPARTITE GRAPHS

A. From unipartite graph modularity to bipartite graph bimodularity

Most authors introduce modularity into bipartite graphs using a probabilistic analogy with the modularity for unipartite graphs. Conversely, we have formally derived bipartite graph modularity from unipartite graph modularity and refer to it as "bimodularity" since it involves both types of nodes in the communities.

In formal terms, a bipartite graph $G = (U, V, E)$ is a graph $G' = (N, E)$ where node set N is the union of two independent sets U and V and moreover the edges only connect pairs of vertices (u, v) where u belongs to U and v belongs to V .

$$N = U \cup V,$$

$$U \cap V = \emptyset,$$

$$E \subseteq U \times V$$

$$\text{Let } r = |U| \text{ and } s = |V|, \text{ then } |N| = n = r + s$$

The unweighted biadjacency matrix of a bipartite graph $G = (U, V, E)$ is a $r \times s$ matrix B in which $B_{i,j} = 1$ iff $(u_i, v_j) \in E$ and $B_{i,j} = 0$ iff $(u_i, v_j) \notin E$.

It must be pointed out that the row margins in B represent the degrees of nodes u_i while the columns' margins represent the degrees of nodes v_j . Conversely, in B^t , the transpose of B , row's margins represent the degrees of nodes v_j and columns' margins represent the degrees of nodes u_i . Let's now define a new matrix as the adjacency matrix A' of G' which is the off-diagonal block square matrix:

$A' = \begin{pmatrix} 0_r & B \\ B^t & 0_s \end{pmatrix}$ where 0_r is an all zero square matrix of order r and 0_s is an all zero square matrix of order s .

Community detection in graphs consists of identifying subsets of densely connected nodes with sparse connections between subsets. Each subset becomes a candidate community. Two main strategies are possible: 1) in the partitions, all communities are distinct; and 2) conversely, overlapping communities may share nodes. As a first step, we will look to identify non-overlapping communities, i.e. graph partitions; then, the second step will involve a generalization to overlapping communities.

Modularity is an indicator often used to measure the quality of graph partitions [29]. First defined for a unipartite graph, several modularity variants have been proposed for bipartite graph partitioning and overlapping communities. These variants will be summarized in a subsequent section. In Equation (2) we propose a new formal definition of modularity for bipartite graphs (which we call herein bimodularity) using the biadjacency matrix. This new definition differs from all modularity definitions previously known for bipartite graphs, which are usually defined *a priori* through probabilistic reasoning. In our specific case, bimodularity is *a posteriori* defined since it has been formally derived from an expression of modularity for unipartite graphs. The corresponding demonstration is detailed in Appendix 1 (Section IX). In the following subsection, we will introduce the notion of bimodularity for bipartite graphs and justify its application for both directed graphs and unipartite graphs, with the aim of unifying the approach for all three types of graphs.

1. Bimodularity for bipartite graphs

Let $G = (U, V, E)$ be a bipartite graph with its biadjacency matrix B and adjacency off-diagonal block matrix A' . Since A' is symmetric, a unipartite graph $G' = (N, E)$ defined by this matrix actually exists. Let's consider Newman's modularity [29] for this graph G' . It is a function Q of both matrix A' and the communities detected in G' :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A'_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (1)$$

where A'_{ij} represents the weight of the edge between i and j , $k_i = \sum_j A'_{ij}$ is the sum of the weights of the edges attached to vertex i , c_i denotes the community to which vertex

i is assigned, the Kronecker's function $\delta(u, v)$ equals 1 if $u = v$ and 0 otherwise and $m = 1/2 \sum_{ij} A'_{ij}$. Hereafter we only consider binary graphs and weights are equal to 0 or 1.

After several transformations we show (see Appendix 1, Section IX) that this modularity can also be written using the biadjacency matrix B of the bipartite graph $G = (U, V, E)$:

$$Q = \frac{1}{m} \sum_{ij} [B_{ij} - \frac{(k_i + k_j)^2}{4m}] \delta(c_i, c_j) \quad (2)$$

where k_i is the margin of row i in B , k_j the margin of column j in B and $m = \sum_{ij} B_{ij} = \frac{1}{2} \times \sum_{ij} A'_{ij} = m$ in (1).

Another interesting formulation to be used is the following (Appendix 1, Section IX):

$$Q = \sum_c [\frac{|e_c|}{m} - (\frac{d_{u|c} + d_{v|c}}{2 \times m})^2] \quad (3)$$

where $|e_c|$ is the number of edges in community c , and $d_{w|c}$ is the degree of node w belonging to c .

Since in the general case B is not symmetric, this definition thus characterizes modularity for bipartite graphs after their extension into new unipartite graphs. It then becomes possible to apply any partitioning algorithm for unipartite graphs to matrix A' and obtain a result where both types of nodes are bound in the same communities, except in the case of singletons (i.e. nodes without edges). To distinguish this definition from unipartite graph modularity and given that it is able to bind both types of nodes, we have called it bimodularity and in Section III C will compare it with other authors' modularity models for bipartite graphs.

2. Bimodularity for *directed* graphs

A directed graphs is a of the form $G^d = (N, E^d)$ where N is a set of nodes and E^d is a set of ordered pairs of nodes belonging to N : $E^d \subseteq N \times N$. From the model in (1) Leicht [21] use probabilistic reasoning 'insights' to derive the following modularity for directed network:

$$Q = \frac{1}{m} \sum_{ij} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta(c_i, c_j) \quad (4)$$

where k_i^{in} and k_j^{out} are the in - and out- degrees of vertices i and j , A is the asymmetric adjacency matrix, and $m = \sum_{ij} A_{ij} = \sum_i k_i^{in} = \sum_i k_j^{out}$. Symmetry is then restored and

spectral optimization applied to extract non-overlapping communities. This model leads to a node partition that does not distinguish between the in and out roles; the nodes are simply clustered within the various communities.

To compare these authors' method to ours, we transformed directed graphs into bipartite graphs (this transformation was also suggested in Guimera's work [17] when applying their method for bipartite networks to directed graphs, as will be seen below). At this point, let's differentiate the nodes' roles into $N \times N$. Along these lines, we duplicate N and consider two identical sets N^{out} and N^{in} . The original directed graph G^d is transformed into a bipartite graph $G = (N^{out}, N^{in}, E)$ in which nodes appear twice depending on their 'out' or 'in' role and moreover the asymmetric adjacency matrix A plays the role of biadjacency matrix B in bipartite graphs. We can now define bimodularity for directed graphs as follows:

$$Q = \frac{1}{m} \sum_{ij} [A_{ij} - \frac{(k_i^{in} + k_j^{out})^2}{4m}] \delta(c_i, c_j) \quad (5)$$

After applying any algorithm for a unipartite graph on the corresponding adjacency matrix A' we obtain a partition where some nodes may belong to the same community twice or instead may appear in two different communities. Each model has its pros and cons. Leicht's model [21] is preferable when seeking a single partition with no role distinction. Our model is attractive when seeking to distinguish between 'in' and 'out' roles, e.g. between producers and customers where anyone can play either role. The brain data example that follows will demonstrate that our model is particularly well suited for analyzing real data.

3. Bimodularity for unipartite graphs

In the above presentation, we introduced bimodularity for bipartite graphs as a formal derivative of unipartite graph modularity. It is dually possible to consider unipartite graphs as bipartite graphs, in defining unipartite graph bimodularity and in extracting communities as if unipartite graphs were bipartite graphs. To proceed, we must consider the original symmetric adjacency matrix A as an asymmetric biadjacency matrix B B (with the same nodes on both dimensions) and build a new adjacency matrix A' using the original adjacency matrix A twice on the off-diagonal, as if the nodes had been cloned. When applying a unipartite graph partitioning algorithm, we then obtain communities in which all nodes appear twice. This method only works if we add to A the unity matrix I (with the same

dimensions as A) before building A' . The first diagonal in A in fact only contains 0s since no loops are generally present in a unipartite graph adjacency matrix. Semantically adding I to A means that all objects will be linked to their respective clones in A' . This is a necessary step in that when extracting communities, the objects must drag their clones into the same communities in order to maintain connectivity. In practice therefore, for unipartite graphs, we build A' with $A + I$.

It may seem futile to perform such a transformation from a unipartite graph to a bipartite one in order to find communities in unipartite graphs given that for computing bipartite graph partitioning, we have already made the extension into unipartite graphs using their (symmetric) adjacency matrix. This transformation is nonetheless worthwhile for several reasons. First, when appearing twice, nodes should be associated with their clones. If the resulting communities do not display this property, i.e. a node's clone lies in another community, then the original matrix is not symmetric and can be considered as the adjacency matrix of a directed graph. This conclusion has been applied to the human brain tractography data clustering, which will be described in the experimental section below.

Conversely, if we are sure that the original adjacency matrix is symmetric, then a result where all nodes are associated with their clones in the same communities would be a good indicator of the quality of the clustering algorithm and moreover provides the opportunity to compare our bipartite graph approach with other unipartite graph strategies. This is also a method we introduced into our experiment (see the karate application below) for the purpose of verifying the validity of results.

Lastly, the most important benefit consists of building overlapping communities and ownership functions for unipartite graphs using the method explained in Section IV below. Although transforming unipartite graphs into bipartite graphs requires more computation, it also provides considerable information, which justifies its application in a variety of contexts.

B. Unifying bipartite and unipartite graph partitioning

The modularity in Equation (1) merely considers unipartite graphs. The bimodularity in Equation (2) considers two types of nodes, with both types in the same community set. We can now show that it is feasible to easily compute partitioned communities in bipartite graphs by using their adjacency off-diagonal matrices and applying unipartite graph partitioning

algorithms.

G' is a unipartite graph; consequently, it is possible to apply its adjacency matrix A' to any algorithm for extracting communities from unipartite graphs. A' is also the off-diagonal adjacency matrix of bipartite graph G . Since bimodularity in Equation (2) has been formally derived from modularity in Equation (1), then computing communities in the unipartite graph G' using symmetric matrix A' and the modularity of Equation (1) is equivalent to computing communities in G using matrix B and bimodularity in Equation 2. As a result, we obtain communities for the corresponding bipartite graph G where both types of nodes are bound.

C. Comparing bimodularity with other modularity models and partitioning algorithms

Several modularity models have been proposed in the literature for unipartite, bipartite, directed weighted and unweighted graphs. This section will compare our bimodularity model with the main proposals from other researchers.

1. Standard probabilistic modularity models for bipartite graphs

Most modularity models for bipartite graphs are derived from Newman’s modularity for unipartite graphs. Authors typically use “ A ” to label their biadjacency matrix, i.e. the matrix whose rows represent one set and columns the other set. We have chosen herein to call this matrix B in order to avoid confusion with the symmetric matrix A for unipartite graphs; hence, we used A' to refer to the corresponding adjacency matrix, which is the off-diagonal block square matrix.

Using such definitions, some authors have probabilistically defined modularity as follows [2][22]:

$$Q = \frac{1}{m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{m} \right] \delta(c_i, c_j) \quad (6)$$

where $m = \sum_{i,j} A_{ij}$, k_i is the margin of row i in A and k_j the margin of column j in A

Differences may exist in the nomenclature from one author to the next, but their definitions remain equivalent. Let’s call this model the standard probabilistic modularity model

for bipartite graphs since it is the most common reference, though we will see below that some authors have slightly adapted this reference.

To compare our bimodularity definition for bipartite graphs with Equation (6) we must first rename the other biadjacency matrix A with our equivalent biadjacency matrix B and replace $m = \sum_{ij} A_{ij}$ by $m = \sum_{ij} B_{ij}$. Once these changes have been introduced, their new definition becomes:

$$Q = \frac{1}{m} \sum_{i,j} [B_{ij} - \frac{k_i k_j}{m}] \delta(c_i, c_j) \quad (7)$$

which is to be compared with our bimodularity formula:

$$Q = \frac{1}{m} \sum_{ij} [B_{ij} - \frac{(k_i + k_j)^2}{4m}] \delta(c_i, c_j) \quad (8)$$

The effects of using either this formula or the other can be observed in the number of communities in each set, as well as in the node distribution by type through modularity optimization. According to our definition, both types of nodes are explicitly bound. If we were to consider the unipartite graph represented by the off-diagonal matrix A' , we can apply any unipartite graph algorithm for detecting communities, and both types of nodes are then regrouped into the same communities (except for singletons). This side effect is not explicit in Equation (5). However since $\delta(c_i, c_j)$ in Equation (5) specifies that the summation is applied to both types of objects belonging to the same community, the side effect is the same: optimizing the standard bipartite graph modularity should yield a partitioning of both types of nodes in the same communities (this analysis is also found in [25] : i.e.: “This definition implicitly indicates that the numbers of communities of both types are equal”). Both modularities should then produce the same results in terms of node type distribution.

As far as the number of communities and node ownership are concerned, it is more difficult to compare the results of both these models, in particular if various algorithms are applied depending on the selected model. For instance, in the Southern Women experiment described below, we found 3 communities when applying Louvain, while [22] found four communities using their original LPAb+ algorithm. These authors however only provided a quantitative evaluation via comparison with other algorithms on computation performance and modularity optimization; in contrast, we have provided qualitative analysis as well, which allows for semantics justification on the partitioning.

2. Bimodularity versus other bipartite graph modularity models

All bipartite graph modularity models analyzed in this section have been derived from the original Newman-Girvan model given in Equation (1).

Guimera [17] introduced an original probabilistic model after considering node co-occurrences of one set ('actors') in the other set ('teams'). Partitioning using simulated annealing is only applied to one set at a time, introducing an important difference with respect to our model, which binds both types of nodes in the same communities. Another valuable contribution by the authors consists of modeling directed graphs as bipartite graphs, in the same way we did in the previous section.

Barber's model [2] is particularly instructive since it starts from the off-diagonal adjacency matrix A to define modularity for bipartite graphs, just like our efforts. This approach however yields a different result when defining modularity. Barber considered the above standard probabilistic model for bipartite graphs and, consequently, did not apply any algorithm to the unipartite graph; instead he applied a complex spectral model, whereby optimization is recursively applied to one set then the other set, thus leading to partitioning where the same number of communities exists for both sets. In our symmetric method, both partitions are computed at the same time using a simple unipartite graph algorithm. In Barber's result, communities of one node type are not directly linked to communities of the other node type; in our case, both node types are bound in communities, with the benefit being that a given node type introduces semantics into regrouping the other node type and *vice versa*.

Murata [25] introduced a new bipartite graph modularity model after critiquing Barber's and Guimera's models for producing equal numbers of communities in both sets. The authors in [25] also compared various modularities and algorithms regarding the issue of modularity optimization applied to a bipartite graph benchmark. In their comparison, three bipartite modularities were found to be the best candidates when at least four communities are present. These three models compared were Barber's, Murata's and Suzuki's, all of which will be analyzed below. It is worthwhile to observe that Barber's symmetric model yields as good results as its two asymmetric competitors. In [22] Murata et al. reintroduced Barber's model along with the implicit aggregation of both sets in the same communities.

With the aim of avoiding equal numbers of communities (symmetric clustering according to the authors), [39] introduced another 'shared clustering' modularity measurement for

bipartite graphs and compared it to Newman’s and Murata’s model. Experiments were limited to the Southern Women’s benchmark, and it will be shown in our experimental section that even with such a benchmark, the superiority of this model has not been proven. Moreover, as explained above, Murata [25] compared this model with other models and could not confirm the fact that asymmetric models are better than symmetric models.

3. *Conclusion on bimodularity versus other bipartite graph modularity models*

After several proposals, it appears that the most widely accepted model for bipartite graph modularity is what we refer to as the ‘standard probabilistic model’, which while not proven has been intuitively proposed by Barber, as inspired by Newman’s model for unipartite graphs. This model implicitly aggregates both sets of nodes in the communities. In our model, we have transformed bipartite graphs into unipartite graphs, and bimodularity has been formally derived from unipartite graph modularity. As a result, we can apply any unipartite graph algorithm and expect partitioning wherever both types of nodes are associated. Far from being a drawback, we will now show that this association is especially attractive in the field of overlapping detection and community analysis.

IV. DETECTION AND ANALYSIS OF COMMUNITY OVERLAPPING

A. Adding semantics to communities

The fact that both types of nodes are bound in their communities yields several important results. First, in considering one type of nodes, a community can be defined by associating a subset of nodes from the other type. In other words, nodes from one set provides semantics for the grouping of nodes from the other set and moreover may qualitatively explain regroupings, as will be seen below. This semantic perspective has not been considered by any of the other authors, a situation due to the fact that in other contributions, either the number of communities differs for both types of nodes (e.g. [24], or else when both types of nodes contain the same number of communities they are not bound in each community [2, 17].

Binding both types of nodes into the same communities yields other pertinent results. For one thing, it is possible to define belonging functions and consequently obtain quanti-

fied overlapping communities. In the following discussion, we will consider three possible belonging functions, which may expose community overlapping in a different light.

B. Probabilistic function

Let's adopt the Southern Women's benchmark, which will be more thoroughly described in Section V below. Applying the Louvain community detection algorithm for unipartite graphs yields a partition where Women and Events are regrouped into three exclusive communities. Let's call these communities c_1 , c_2 and c_3 . Now, let's suppose the fictitious case in which woman w_1 participated in events e_1 , e_2 , e_3 and e_4 . furthermore, w_1 , e_1 and e_2 are classified in c_1 , while e_3 is classified in c_2 and e_4 is classified in c_3 . We can then define a probability function as follows:

$$P(u_i \in c) = \frac{1}{k_i} \sum_j B_{ij} \delta(c_j) \quad (9)$$

where c is a community, $k_i = \sum_j B_{ij}$ and $\delta(c_j) = 1$ if $v_j \in c$ or $\delta(c_j) = 0$ if $v_j \notin c$

In $P(u_i \in c)$ the numerator includes all edges linking u_i to properties $v_j \in c$ and the denominator contains all edges linking u_i to all other nodes. With this function in the present example the probability of w_1 being classified in community c_1 equals $\frac{2}{4}$, and her probabilities of being classified in c_2 and in c_3 are $\frac{1}{4}$ each. The probability a node belongs to a given community is the percentage of its links to this community as a proportion of the total number of links to all communities. In other words, the greater the proportion of links to a given community, the higher the expectation of belonging to this community.

C. Legitimacy function

It is possible to add more semantics in order to decide which community a given node should join. The legitimacy function serves to measure the node involvement in a community from the community's perspective. The more strongly a node is linked to other nodes in a community, the greater its legitimacy to belong to the particular community. In the Southern Women's example, let's assume that after partitioning, c_1 contains 7 events, c_2 5 events and c_3 2 events (which is actually the case in the experiment presented below). Then, w_1 would have a $\frac{2}{7}$ legitimacy for c_1 , $\frac{1}{5}$ for c_2 and $\frac{1}{2}$ for c_3 . The legitimacy function can thus be formalized as follows:

$$L(u_i \in c) = \frac{\sum_j B_{ij} \delta(c_j)}{|\{v \in c\}|} \quad (10)$$

where c is a community, $\delta(c_j) = 1$ if $v_j \in c$ or $\delta(c_j) = 0$ if $v_j \notin c$

The numerator in this expression is the same as the probabilistic function numerator. Only the denominator is different.

D. Reassignment Modularity function

Reassigning node w from C_1 to C_2 either increases or decreases the modularity defined in Equation(1). Such a change is referred to as Reassignment Modularity ($RM_{w:C_1 \rightarrow C_2}$).

Let w be a node u or v . If w is withdrawn from C_1 and reassigned to C_2 , then we can define $RM_{w:C_1 \rightarrow C_2} = Q_{w \in C_2} - Q_{w \in C_1}$

where Q is the modularity value in

$$Q = \sum_c \left[\frac{|e_c|}{m} - \left(\frac{d_{u|c} + d_{v|c}}{2 \times m} \right)^2 \right] \quad (11)$$

Let $l_{w|i} = l_{w,w'|w' \in C_i}$ be the number of edges between w and w' where $w' \in C_i$,

Let d_w be the degree of w , $|e_i|$ the number of edges in C_i and $d_{C_i} = d_{u|C_i} + d_{v|C_i}$

then

$$Q_{w \in C_2} - Q_{w \in C_1} = \left[\frac{1}{m}(|e_1| - l_{w|1}) + \frac{1}{m}(|e_2| + l_{w|2}) - \left(\frac{(d_{C_1} - d_w)^2}{(2m)^2} + \frac{(d_{C_2} + d_w)^2}{(2m)^2} \right) \right] - \left[\frac{1}{m}|e_1| - \frac{(d_{C_1})^2}{(2m)^2} + \frac{1}{m}|e_2| - \frac{(d_{C_2})^2}{(2m)^2} \right]$$

and

$$RM_{w:C_1 \rightarrow C_2} = \frac{1}{m}(l_{w|2} - l_{w|1}) - \frac{2}{(2m)^2}[d_w^2 + d_w(d_{C_2} - d_{C_1})] \quad (12)$$

This equation can be partly validated if after withdrawing w from C_1 we put it back into C_1 : considering that C_2 is in fact C_1 without w , we get $d_{C_2} = d_{C_1} - d_w$. Then replacing d_{C_2} in equation (20) by its value yields $RM_{w:C_1 \rightarrow C_1} = 0$.

A second validation can be performed with Equation (IX B) in [40]. Although the authors' demonstration is limited, it can still be noticed that their final formula resembles ours with a slight difference (i.e. division by 2 in their case) due to their definition of modularity for overlapping communities. Moreover, in arguing that the right part of their equation is not meaningful for large graphs, the authors only considered $dEQ = \frac{l_2 - l_1}{2m}$ which is the equivalent of $\frac{1}{m}(l_{w|2} - l_{w|1})$ in our Reassignment Modularity definition. In our case, we did not limit reassignment to large graphs and moreover kept the whole value in Equation (20).

E. Other functions

Other assignment criteria may be explored, such as cardinality of communities: when hesitating in a tie, an individual may prefer joining a small community or, conversely, a large one. These criteria may be considered in practical applications but will not be taken into account within the scope of the present paper. Among the measures presented in this section, two will be used to display our results, namely the Legitimacy function and the Reassignment Modularity function. These two functions are of tremendous value in inferring semantic analyses on overlapping results.

V. VISUALIZATION

Visualization is of utmost importance since it provides a means for analyzing and interpreting the communities extracted from graphs. In most cases, unipartite graphs are displayed using vertices and edges. Though these graphs in a large and non-planar form quickly become entangled, it is still possible to observe a certain structure by the use of graph drawing techniques [3]. Visualizing partitioning is simple on unipartite graphs through vertex coloring, provided the number of communities remains limited, such as in [4]. With regard to community overlapping, the visualization step is much more complicated (even for unipartite graphs) since vertices may belong to several communities and, consequently, cannot be easily assigned several colors. One possible solution would be to consider the community assignment as a bipartite graph with one set of nodes being the graph vertices and the other set being the communities and edges linking vertices to their communities. Bipartite graphs can be represented like any other graphs or as two-layer graphs, with both sets being organized in lines either horizontally or vertically. When community assignment is represented with edges however, the original edges between vertices cannot be shown on the same graph without creating confusion. It is therefore more appropriate to separately visualize the overlap through a biadjacency matrix, whose rows represent communities and columns the original vertices. From this perspective, each community is depicted with its core members and each vertex with the various communities where it belongs. Other visualization techniques have been studied, particularly when considering hypergraphs on which communities are the hyperedges and vertices the nodes. For example, Euler diagrams have

been explored in [35] and [37], Galois lattices in [5, 18], and linesets in [1], they are limited however by the hypergraph size, with no clear representation emerging for medium-sized hypergraphs. For this reason, we have only considered the biadjacency visualization strategy in the Karate club experiment conducted below, which proves to be valuable for our visual analysis of community overlapping. Moreover, this representation technique offers the possibility of indicating ownership functions, such as legitimacy (see Section IV C), which is not possible with other representations. When unipartite graphs are not excessively large, it becomes important to retain all original information; it is also possible to show the original unipartite adjacency matrices and then add community overlapping. This possibility will not be discussed within the scope of the present paper, which is solely dedicated to studying the community overlapping visualization and does not consider the details of vertex relations. Conversely, when the graph is too large, it is no longer possible to even show community details, in which case it is better to consider a square stochastic block matrix that only displays communities with their relationship degrees (see [11] and [16]).

The visualization of community overlapping is even more complicated for bipartite graphs since two of these graphs need to be represented: the original graph, and the computed community overlap structure. As explained above, two main strategies are available to visualize bipartite graphs: a bilayer graph, and the biadjacency matrix. In limiting our objective to representing partitioning, any symbolic representation, such as node shape or coloring in the bilayer graph or even in the original graph, proves sufficient (see for instance Figure 2 in [2] or Figure 6 in [39]). Visualizing community overlap requires alternative strategies. Along these lines, we have explored two matrix representations. A detailed representation, not presented within the scope of this paper, is based on the biadjacency matrix of the original bipartite graph. The other bipartite representation, in which rows are communities and columns people, will be provided below. Two dual representations may coexist for a given situation, one for the first set the other for the second, with both sets sharing the same communities. In the Southern Women’s experiment described below, we analyze such a dual representation while retaining the original bilayer bipartite graph. The two other experiments introduced in the next section will be limited to representing one set of vertices according to the semantic focus.

VI. EXPERIMENTATION

This section will consider several benchmarks from various sources. We begin by applying our method to two simple graphs: the so-called "karate club" unipartite graph from [42] shows friendship relations between members of a karate sport club; and the "Southern Women" bipartite graph depicts relations between southern American women participating in several events. Our method is then applied to a medium-sized dataset extracted from a real-world situation. For this purpose, we consider a bipartite graph (people tagged on photos) drawn from a student's "Facebook" account containing an average number of photos and people. Lastly, this same method will be applied to human brain data in order to derive dependencies between several areas in the brain.

A. Karate club

The karate club graph [42] is a well-known benchmark showing friendship relations between members of a karate club; it is a unipartite graph on which many partitioning algorithms have been experimented. Consequently, this set-up makes it possible not only to verify that our method for bipartite graphs when applied to unipartite graphs meets expected results, but also to assess the additional knowledge extracted from overlapping.

We began by directly applying the Louvain algorithm to the original unipartite graph, represented by its adjacency matrix A . which yielded four separate communities (as shown in 2). These are the same communities extracted by other authors, e.g. [29]. During a second experiment, we considered that the adjacency matrix A is in fact a biadjacency matrix B which is representative of a bipartite graph whose corresponding objects are the club members and whose properties are also club members. An edge exists in the bipartite graph between a club member-object and a club member-property provided an edge is present between the two club members in the original unipartite graph. The new A' adjacency matrix is $A' = \begin{bmatrix} O_r & B \\ B^t & O_s \end{bmatrix}$, where $B = A + I$. and where I is the identity matrix (as explained in section III A 3). We once again apply the Louvain algorithm to A' .

Results. As expected, these same four communities identified in the unipartite graph have been extracted from the bipartite graph, with the same individuals appearing twice in each community (see Figure 2). This initial result confirms the absence of bias when

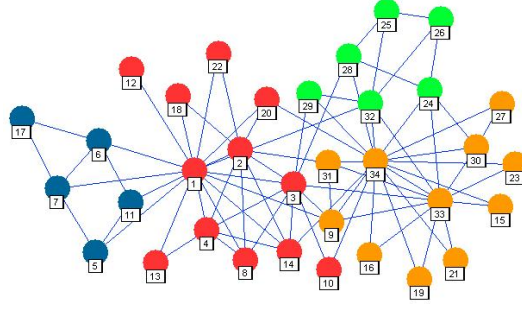


Figure 1: Karate club graph with partitioned communities

(see a color version at this web address : www.lgi2p.ema.fr/plantie/PR/FIGURE-1.jpg)

transforming a unipartite graph into a bipartite one. The second result is more pertinent because it reveals an overlap between communities when considering legitimacy values. If we were to consider just the cell colorings in the figure, an overlap would be observable whenever at least one node from a community is linked to other nodes in another community. The legitimacy values that indicate the involvement of each node in each community offer an effective tool for identifying and analyzing new features. Some slight differences have been noted in works by other authors: for example, in page 2, Porter [34] placed node number 10 in the second community. In our case, this node has been placed in the first community, though the legitimacy value suggests that it should have been placed in the second one, in which case the situation would be reversed in the second community and node 10 would have a legitimacy value that alters its placement in the first community. Node 10 is thus in a hesitation mode between the two communities.

To the best of our knowledge, this experiment represents the first time Karate communities are shown as separate and overlapping. Partitioning provides a practical way to observe communities; however, overlapping reveals the extent to which partitioning reduces the amount of initial information. With our method for example, it can be seen that some nodes actually straddle several communities, e.g. node 10 in our experiment.

B. Southern Women

This benchmark has been studied by most authors interested in checking their partitioning algorithm for bipartite graphs. The goal here is to partition, into various groups, 18 women who attended 14 social events according to their level of participation in these events. In

node N°		1	2	3	4	8	10	12	13	14	18	20	22	9	15	16	19	21	23	27	30	31	33	34	24	25	26	28	29	32	5	6	7	11	17	
Unipartite community N°		1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4
overlapping community + legitimacy	1	10/12	8/12	6/12	6/12	4/12	1/12	1/12	2/12	4/12	2/12	2/12	2/12	2/12	0	0	0	0	0	0	0	0	1/12	1/12	3/12	0	0	0	1/12	1/12	1/12	1/12	1/12	1/12	1/12	0
	2	1/11	1/11	2/11	0	0	1/11	0	0	1/11	0	1/11	0	3/11	2/11	2/11	2/11	2/11	2/11	2/11	2/11	3/11	3/11	9/11	10/11	3/11	0	0	1/11	1/11	2/11	0	0	0	0	0
	3	1/6	0	2/6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1/6	0	2/6	4/6	2/6	3/6	3/6	2/6	1/6	3/6	0	0	0	0	0
	4	4/5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2/5	3/5	3/5	2/5	2/5

Figure 2: Karate club communities and modularity measures

(see a color version at this web adress : www.lgi2p.ema.fr/plantie/PR/FIGURE-2.jpg)

his well-known cross-sectional study, [13] compared results from 21 authors, most of whom identified two groups.

Results. In Figure 3, the bipartite graph is depicted as a bilayer graph in the middle with women at the top and events below; moreover, the edges between women and events represent woman-event participations. Three clusters with associated women and events have been found and eventually shown with black, dark grey and light grey colorings. This result is more accurate than the majority of results presented in [13]; only one author found three female communities. Beyond mere partitioning, Figure 3 presents overlapping communities using two overlapping functions, namely legitimacy and reassignment modularity (RM). Legitimacy overlapping and RM for women are placed just above female partitioning; for events, both are symmetrically shown below event partitioning. As expected, reassignment in the same community produces a zero RM value. The best values for legitimacy and RM have been underscored. Only the values of woman 8 and event 8 indicate that they could have been in another community. This is the outcome of early assignment during the first Louvain phase for entities with equal or nearly equal probabilities across several communities. It can be observed in [13] that woman 8's community is also debated by several authors; our results appear to be particularly pertinent in terms of both partitioning and overlapping.

One possible criticism of our result may stem from the fact that women and events are correlated, which may cause bias, such as in the number of communities. When comparing our results to those of other authors however, the merging of our blue and yellow communities produces their corresponding second community. In their trial designed to obtain a varying number of communities in both sets, Suzuki [39] found a large number of singletons. Their

3		-41,54		-41,54					6,44	-2,90	-5,68	-16,92	-37,37	-36,55	-13,38	-15,84	0,00	0,00	0,00
2	RM	-54,66	-54,41	-54,66	-54,41		-20,70	-20,70	1,51	-9,47	0,00	0,00	0,00	0,00	0,00	0,00	-3,53		
1		0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-15,02			-32,90	-32,07	-27,52			
3		1/2		1/2					1/2	1/2	1/2	1/2	1/2	1/2	2/2	1/2	1/2	2/2	2/2
2	Legitimity	1/5	1/5	1/5	1/5		1/5	1/5	1/5	1/5	2/5	3/5	5/5	5/5	4/5	3/5	1/5		
1		6/7	6/7	6/7	6/7	4/7	3/7	3/7	1/7	2/7	1/7			1/7	2/7	1/7			
Women		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18
Events		E1	E2	E3	E4	E5	E6	E7			E8	E10	E12	E13	E14			E9	E11
1		3/9	3/9	6/9	4/9	8/9	7/9	6/9			8/9							4/9	
2	Legitimity						1/6	4/6			5/6	5/6	6/6	3/6	3/6			5/6	2/6
3											1/3							3/3	2/3
1		0,00	0,00	0,00	0,00	0,00	0,00	0,00			11,61							-25,88	
2	RM						-65,90	-21,84			0,00	0,00	0,00	0,00	0,00			-6,31	-7,57
3											-34,34							0,00	0,00

Figure 3: Women Events communities and modularity measures

(see a color version at this web adress : www.lgi2p.ema.fr/plantie/PR/FIGURE-3.jpg)

results were far from those presented in [13], while ours were compatible and more highly detailed.

In conclusion, results on the Southern Women’s benchmark are particularly relevant. Moreover, our visualization enables observing community partitioning, overlapping and possible assignment contradictions. The application of reassignment for better modularity optimization will be tested in a subsequent work.

C. Facebook account

Three Facebook photo files were downloaded from various Facebook (FB) accounts. All these files were extracted with the consent of their owners, none of whom were members of the research team. A person was considered to be linked to a photo if he/she had been tagged in the photo producing a bipartite graph. We evaluated data from FB photo tags and not friendship relations. Community extraction using our method reveals some common features among the datasets. These features are shown in Figure 4 for one FB photo file, in which 274 people could be identified in a total of 644 photos.

Results. Communities are seldom overlapping, which supports the notion that the photos were taken at different times in the owner’s life (this is to be confirmed in a forthcoming study). When the owner was asked to comment on the communities, two main observations were submitted. The various groups of people were indeed consistent, yet with one exception. The owner was associated in the partition with a group she had met on only a few occasions



Figure 4: Facebook account communities and modularity measures

(see a color version at this web adress : www.lgi2p.ema.fr/plantie/PR/FIGURE-4.jpg)

and not associated with other groups of close friends. An analysis of the results provided a good explanation, which is partially displayed in Figure 4. From this view, the FB account owner is in the first community on the left, yet she is also present in most of the other communities (see grey color levels in the first column). Although at first glance it might be assumed that she is not part of other communities, our visualization indicates that such is not the case. She is present in most communities, even though she is mainly identified in the first one. Three types of photos can be distinguished in this first community. More than 200 photos only contain the owner's tag, plus a few photos with unique tags of another community member; for every other person, at least one photo tags him/her with the owner. This first community has in fact been built from the first group with photos of unique owner's tags associated with the owner. The owner's tag thus encompasses photos containing two people, one of whom is the owner. It turns out that this group is predominantly the owner's group.

In conclusion, partitioning only the bipartite graph would have produced a major pitfall: the owner would have been isolated in a community that is not his/her top preference. With our method, merging partitioning and overlapping exposes better multiple regroupings with broader affinities. Other communities also showed high consistency when considering the photos: each community was associated with some particular event responsible for gathering a group of the FB account owner's friends.

D. Brain Data

Our method was initially designed for human community detection and analysis. In this experiment, we have demonstrated how it can be applied to other data analysis techniques as well. The brain dataset was collected on a single patient by a research team affiliated with the "Human Connectome" project working on brain tractography techniques [6]. These techniques use Magnetic Resonance Imaging (MRI) and Diffusion Tensor Imaging (DTI) to explore white matter tracks between brain regions. Probabilistic tractography produces 'connectivity' matrices between Regions Of Interest (ROI) in the brain. For the case we studied, 'seed' ROIs were located in the occipital lobe and 'target' ROIs throughout the entire brain. The goal here was to detect possible brain areas in the occipital lobe through ROI clustering on the basis of similar track behavior. In [6], the research team used Spectral Clustering (SC) to combine ROIs. It is interesting to note that SC is one of numerous techniques that have traditionally been applied in social community detection, e.g. by Bonacich on the Southern Women's benchmark [13]. SC results are limited to community partitioning (though in theory overlapping could also be computed). The goal was to experiment with our method and produce both partitioning and overlapping analyses of brain areas.

The original matrix contained 1,914 rows and 374 columns, with cells denoting the probabilities of linkage between ROIs. We considered this matrix as a bipartite graph biadjacency matrix with weighted values and then applied our community detection method. Figure 6 presents the results of ROI community partitioning and overlapping. Each grey level color in the first row is associated with a community that gathers several ROIs. Each ROI is represented by a column that indicates its belonging to the other communities. When a cell is highlighted with a grey color, a nonzero overlapping value exists for both this ROI and the corresponding community (with community numbers being plotted on the left-hand side of the figure). This value has been computed with the legitimacy function, which has been extended to the weighted edges, i.e. the weighted sum of values from cerebral hemisphere zones (ELF) within the selected community. Each community is associated with a threshold value corresponding to the maximum weighted legitimacy above which the community would lose a full member. For each community, this threshold value is automatically computed in order to include all ROI members of the community.

Results. We found 7 communities when neurologists selected 8 clusters with SP and after

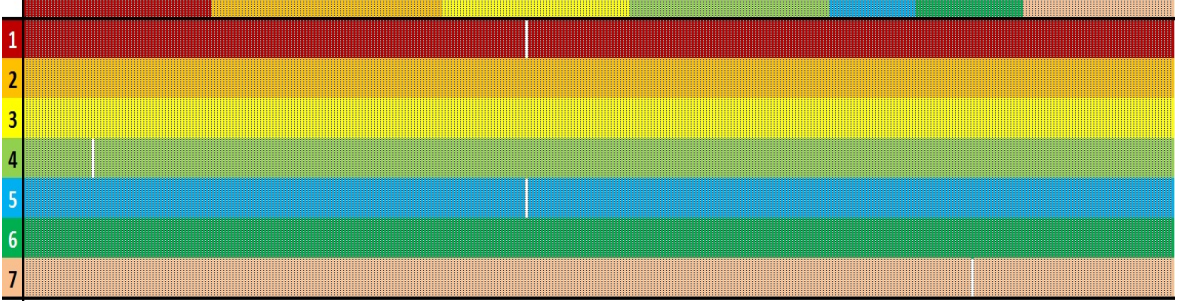


Figure 5: Brain data communities and modularity measures

(see a color version at this web adress : www.lgi2p.ema.fr/plantie/PR/FIGURE-5.jpg)

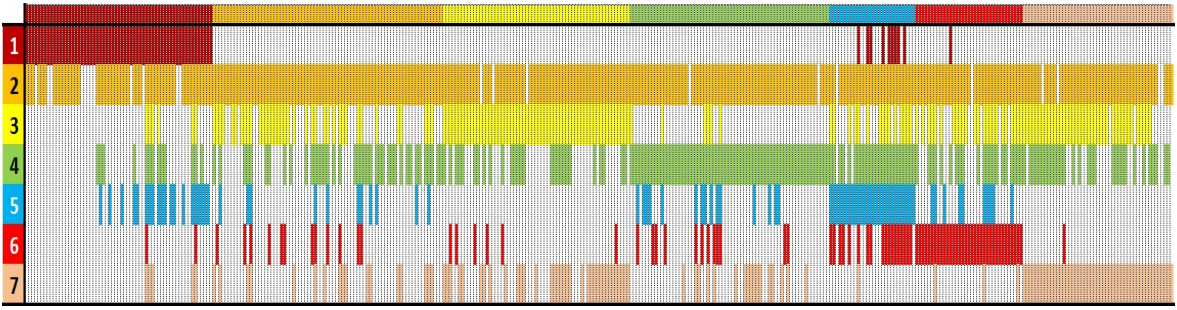


Figure 6: Brain data communities and modularity measures

(see a color version at this web adress : www.lgi2p.ema.fr/plantie/PR/FIGURE-6.jpg)

choosing the most significant eigenvectors on a scree test. Let's observe that two communities overlap heavily on all others, which thus overlap to a lesser extent. If we were to lower all thresholds to a zero value, then the overlapping would be maximized (see Figure 5). Figure 6 confirms the strong interest in this set-up that simultaneously exhibits overlapping and non-overlapping data. These results have been taken into account by a team of neurological researchers as different observations recorded on brain parcellation.

VII. DISCUSSION ET CONCLUSION

In this paper, we have demonstrated the feasibility of unifying bipartite graphs, directed graphs and unipartite graphs after: considering them as bipartite graphs with their bi-adjacency matrix, then building the corresponding unipartite graph with the off-diagonal adjacency matrix, and finally preparing for community building through a unipartite graph partitioning algorithm. Along these lines, we formally derived a bipartite graph modularity

model from the standard unipartite modularity model. It was then proven that any unipartite graph partitioning algorithm aimed at optimizing the standard unipartite modularity model will lead to a bipartite graph partitioning, wherein both types of nodes are bound in the communities. In the special case of directed graphs, nodes appear twice in potentially different communities depending on their roles; for unipartite graphs, nodes are cloned and appear with their clones in the same communities.

Any approach that produces communities from bipartite graphs associating both types of nodes has been qualified as a 'symmetric' method in [39]. Such approaches have been criticized by some authors, who argue that the number of communities in both types of nodes is often skewed [24, 39]. Though this point of view remains defensible, we have shown that most authors ultimately introduced a standard probabilistic model for bipartite graphs that implicitly associates both types of nodes. Moreover, all the experiments we identified in other papers actually present results that appear to be more accurate with symmetric models.

Our approach is not limited to unifying different types of graphs; we also introduced the possibility of unifying, into just a single view, the partitioning and overlapping communities. This development is possible thanks to associating both types of nodes in the communities. Moreover, overlapping can be characterized through several functions presenting different semantic meanings. For instance, it is possible to identify those nodes that define the community cores, i.e. those who belong exclusively to just one community and, conversely, those who serve as bridges between different communities. Some of these functions even create the possibility to compute reassignment values, which may then be used for fine-tuning the greedy partitioning algorithms.

Practically speaking, when applying our method to various benchmarks and datasets, we are able to extract meaningful communities and display surprising overlapping properties. Other authors' models limit the goal to identifying communities. We extend far beyond this point and provide tools for analyzing and interpreting results. We can understand how some entities may be hesitating between various community assignments, and moreover it is possible to measure this 'fuzziness'. Some entities may even be reassigned in the aftermath in pursuit of improved modularity optimization.

Lastly, we introduced an essential result after experimenting on real brain datasets, supplied by a research team from the Connectome project. Many traditional data analysis

techniques have often been implemented by authors for the purpose of community detection in unipartite or bipartite graphs, e.g. Spectral Clustering or hierarchical clustering. These methods however require specifying a number of clusters or setting a threshold. Recent community detection algorithms, based on modularity detection, do not require any such subjective orientation. We have applied algorithms of this type (e.g. Louvain) so as to work with clustering in data analysis and compare it to traditional data analysis techniques. The results are very similar for both approaches, yet two main differences are noteworthy. First, community grouping is solely dependent on the original data and does not require the *a priori* choice of eigenvalues. Second, we were easily able to provide both partitioning and overlapping communities with ownership functions. This result is of particular interest when dealing with brain data in which community borders are not clear cut.

Acknowledgements

The authors would like to thank the Connectome research team, as part of the “CAFO” project (ANR-09-RPDOC-004-01 project) as well as the CRICM UPMC U975/UMRS 975/UMR 7225 research group for providing the original brain dataset.

VIII. REFERENCES

-
- [1] Basak Alper, Nathalie Riche, Gonzalo Ramos, and Mary Czerwinski. Design Study of LineSets, a Novel Set Visualization Technique, 2011.
 - [2] Michael Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):1–9, 2007.
 - [3] Battista G., Eades, Tamassia, and Tollis. *Graph drawing. Algorithms for the visualisation of graphs*. Prentice Hall, 1999.
 - [4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

- [5] Ulrik Brandes, Sabine Cornelsen, Barbara Pampel, and Arnaud Sallaberry. Path-Based Supports for Hypergraphs. *Order A Journal On The Theory Of Ordered Sets And Its Applications*, pages 1–14, 2010.
- [6] Marco Catani and Michel Thiebaut de Schotten. *Atlas of human brain connections*. Oxford University Press, 2012, 2012.
- [7] Abhijnan Chakraborty, Saptarshi Ghosh, and Niloy Ganguly. Detecting overlapping communities in folksonomies. In *Proceedings of the 23rd ACM conference on Hypertext and social media HT 12*, page 213. ACM Press, 2012.
- [8] Ernesto Estrada and Juan A Rodriguez-Velazquez. Complex Networks as Hypergraphs. *Systems Research*, page 16, 2005.
- [9] T S Evans. Clique Graphs and Overlapping Communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(12):23, 2010.
- [10] T S Evans and R Lambiotte. Line Graphs, Link Partitions and Overlapping Communities. *Physical Review E*, 80(1):9, 2009.
- [11] Katherine Faust and Stanley Wasserman. Blockmodels: Interpretation and evaluation. *Social Networks*, 14(1-2):5–61, 1992.
- [12] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):103, June 2009.
- [13] Linton C. Freeman. Finding social groups: A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis. The National Academies*, pages 39–97. Press, 2003.
- [14] M. Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.
- [15] Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2009.
- [16] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52):22073–8, December 2009.
- [17] Roger Guimerà, Marta Sales-Pardo, and Luís Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3), September 2007.
- [18] Jeffrey Johnson. Hypernetworks for reconstructing the dynamics of multilevel systems. *Net-*

- works*, 2(September):25–29, 2006.
- [19] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, March 2009.
 - [20] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. *4th Workshop on Social Network Mining and Analysis SNAKDD10*, 10:10, 2010.
 - [21] E A Leicht and M E J Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, 2007.
 - [22] Liu Xin and Murata Tsuyoshi. An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14(4):408–415, 2010.
 - [23] Michel Plantié and Michel Crampes. Mining social networks and their visual semantics from social photos. *International Journal of Computer science & Applications*, VIII(II):102–117, 2011.
 - [24] Tsuyoshi Murata. Modularities for bipartite networks. *Proceedings of the 20th ACM conference on Hypertext and hypermedia HT 09*, 90(6):245–250, 2009.
 - [25] Tsuyoshi Murata and Tomoyuki Ikeya. A new modularity for detecting one-to-many correspondence of communities in bipartite networks. In *Advances in Complex Systems*, volume 13, pages 19–31. World Scientific Publishing Company, February 2010.
 - [26] Neubauer Nicolas and Obermayer Klaus. Towards Community Detection in k-Partite k-Uniform Hypergraphs. In *Proceedings NIPS 2009*
 - [27] Mark Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), June 2004.
 - [28] Mark Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 74(3 Pt 2):036104, 2006.
 - [29] Mark Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), February 2004.
 - [30] Mark Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 68(3 Pt 2):036122, 2003.

- [31] Andreas Noack and Randolph Rotta. Multi-level algorithms for modularity clustering. page 12, December 2008.
- [32] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8, June 2005.
- [33] S Papadopoulos, Y Kompatsiaris, A Vakali, and P Spyridonos. Community detection in Social Media. *Data Mining and Knowledge Discovery*, (June):1–40, 2011.
- [34] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in Networks, 2009.
- [35] Nathalie Henry Riche and Tim Dwyer. Untangling euler diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1090–1099, 2010.
- [36] Camille Roth and Paul Bourguine. Epistemic Communities: Description and Hierarchic Categorization. *Mathematical Population Studies: An International Journal of Mathematical Demography*, 12(2):107–130, 2005.
- [37] Paolo Simonetto, David Auber, and Daniel Archambault. Fully Automatic Visualisation of Overlapping Sets. *Symposium A Quarterly Journal In Modern Foreign Literatures*, 28(3):967–974, 2009.
- [38] Sune Lehmann, Martin Schwartz, Lars Kai Hansen. Biclique communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78(1 Pt 2), 2008.
- [39] Kenta Suzuki and Ken Wakita. Extracting Multi-facet Community Structure from Bipartite Networks. *2009 International Conference on Computational Science and Engineering*, 4:312–319, 2009.
- [40] Zhihao Wu, Youfang Lin, Huaiyu Wan, Shengfeng Tian, and Keyun Hu. Efficient overlapping community detection in huge real-world networks. *Physica A: Statistical Mechanics and its Applications*, 391(7):2475 – 2490, 2012.
- [41] Bo Yang, Dayou Liu, Jiming Liu, and Borko Furht. *Discovering communities from Social Networks: Methodologies and Applications*. Springer US, Boston, MA, 2010.
- [42] W W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

IX. ANNEX 1

A. From modularity to bimodularity

In this Appendix, we will provide full details of the demonstration that yielded Equation (2)

For the sake of convenience, let's use the definition of unipartite graph modularity offered in Newman [21]. It is a function Q of matrix A' and the communities detected in G [29]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A'_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (13)$$

where A'_{ij} denotes the weight of the edge between i and j , $k_i = \sum_j A'_{ij}$ is the sum of the weights of edges attached to vertex i , c_i is the community to which vertex i has been assigned, the Kronecker's function $\delta(u, v)$ equals 1 if $u = v$ and 0 otherwise and $m = 1/2 \sum_{ij} A'_{ij}$. Hereafter, we will only consider binary graphs and weights that are equal to 0 or 1.

In our particular case (i.e. where A' is the off-diagonal block adjacency matrix of a bipartite graph), we apply the following transformations:

Let's rename i_1 as index i when $1 \leq i \leq r$ and i_2 when $r < i \leq r + s$. Conversely, let's rename j_1 the index j when $1 \leq j \leq r$ and j_2 when $r < j \leq r + s$.

To avoid confusion between the A' 's indices and B 's indices let's rename B indices i_b and $j_b : 1 \leq i_b \leq r$ and $1 \leq j_b \leq s$ (see a representation of A matrix below (Equation 14))

$$A' = \begin{array}{c|c|c|c|c} A' \text{ indexes} & & & & \\ \downarrow \rightarrow & \dots j_1 \dots & \dots j_2 \dots & & \\ \hline \dots & & & \dots & \\ i_1 & O_r & B & i_b & r \text{ rows} \\ \dots & & & \dots & \\ \hline \dots & & & \dots & \\ i_2 & B^t & O_s & j_b & s \text{ rows} \\ \dots & & & \dots & \\ \hline & \dots i_b \dots & \dots j_b \dots & \leftarrow \uparrow & \\ & & & B \text{ indexes} & \\ \hline & r \text{ columns} & s \text{ columns} & & \end{array} \quad (14)$$

Let's call k_{i_b} the margin of row i_b in B and k_{j_b} the margin of column j_b in B .

$$k_{i_b} = \sum_{j_b} B_{i_b j_b} = \sum_{j_2} A'_{i_1 j_2} = \sum_{i_2} A'_{i_2 j_1}, \text{ where } i_b = i_1 = j_1 \quad (15)$$

$$k_{j_b} = \sum_{i_b} B_{i_b j_b} = \sum_{i_1} A'_{i_1 j_2} = \sum_{j_1} A'_{i_2 j_1}, \text{ where } j_b = i_2 - r = j_2 - r \quad (16)$$

k_{i_b} is the degree of node u_{i_b}

k_{j_b} is the degree of node v_{j_b}

Let's define $k_{i/j_1} = \sum_{j_1} A'_{i j_1}$ and $k_{i/j_2} = \sum_{j_2} A'_{i j_2}$

Conversely : $k_{j/i_1} = \sum_{i_1} A'_{j i_1}$ and $k_{j/i_2} = \sum_{i_2} A'_{j i_2}$

Hence :

$$k_i = \sum_j A'_{ij} = k_{i/j_1} + k_{i/j_2}$$

$$k_j = \sum_i A'_{ij} = k_{j/i_1} + k_{j/i_2}$$

By taking into account the structure and properties of A in (15) and (16) for the indices we derive the following properties :

k_{i/j_1} has non-zero values only for $i = i_2$, with k_{j_b} the degree of node v_{j_b} :

$$k_{i/j_1} = k_{i_2/j_1} = \sum_{j_1} A'_{i_2 j_1} = \sum_{i_1} A'_{i_1 j_2} = k_{j_2/i_1} = k_{j_b} \quad (17)$$

k_{i/j_2} has non-zero values only for $i = i_1$, with k_{i_b} the degree of node u_{i_b} :

$$k_{i/j_2} = k_{i_1/j_2} = \sum_{j_2} A'_{i_1 j_2} = \sum_{i_2} A'_{i_2 j_1} = k_{j_1/i_2} = k_{i_b} \quad (18)$$

Moreover and more directly:

k_{j/i_1} offers values only for $j = j_2$: $k_{j/i_1} = k_{j_2/i_1} = k_{i_2/j_1} = k_{j_b}$, the degree of node v_{j_b} .

k_{j/i_2} offers values only for $j = j_1$: $k_{j/i_2} = k_{j_1/i_2} = k_{i_1/j_2} = k_{i_b}$, the degree of node u_{i_b} .

B. Analysing second part of Q in (13)

Using these properties of matrix A' , it is now possible to analyse $\sum_{ij} k_i k_j$. in equation (1).

Next, by developing k_i and k_j in A' we obtain:

$$\sum_{ij} k_i k_j = \sum_{ij} (k_{i/j_1} + k_{i/j_2})(k_{j/i_1} + k_{j/i_2})$$

$$\begin{aligned}
&= \sum_{ij} k_{i/j_1} k_{j/i_1} + \sum_{ij} k_{i/j_2} k_{j/i_2} + \sum_{ij} k_{i/j_1} k_{j/i_2} + \sum_{ij} k_{i/j_2} k_{j/i_1} \\
&= \sum_{i_2 j_2} k_{i_2/j_1} k_{j_2/i_1} + \sum_{i_1 j_1} k_{i_1/j_2} k_{j_1/i_2} + \sum_{i_2 j_1} k_{i_2/j_1} k_{j_1/i_2} + \sum_{i_1 j_2} k_{i_1/j_2} k_{j_2/i_1} \quad (19)
\end{aligned}$$

Let's note that $\sum_{ij} k_{i/.} k_{j/.} = \sum_i k_{i/.} \sum_j k_{j/.}$ where the dot may take any value in i_1, i_2, j_1, j_2

Let c be a community, in equation (1) summations $\sum_{ij} k_i k_j$ on indices i and j may only be applied under the condition $\delta(c_i, c_j) = 1$. Where an edge is present between two nodes u and v belonging to c : $\delta(c_i, c_j) = 1$ and $\delta(c_j, c_i) = 1$. Consequently for each row i representing a node belonging to c , a corresponding column j represents this same node belonging to c and *vice versa*.

From (17), (18), property (5) and the above observation:

$$\begin{aligned}
\sum_{ij} k_{i/j_1} k_{j/i_1} \delta(c_i, c_j) &= \sum_i k_{i/j_1} \sum_j k_{j/i_1} \delta(c_i, c_j) = \sum_{i_2} k_{i_2/j_1} \sum_{j_2} k_{j_2/i_1} \delta(c_{i_2}, c_{j_2}) = \\
\sum_{j_b} k_{j_b} \sum_{j_b} k_{j_b} &= [\sum_{j_b} k_{j_b}]^2 \\
\sum_{ij} k_{i/j_2} k_{j/i_2} \delta(c_i, c_j) &= \sum_i k_{i/j_2} \sum_j k_{j/i_2} \delta(c_i, c_j) = \sum_{i_1} k_{i_1/j_2} \sum_{j_1} k_{j_1/i_2} \delta(c_{i_1}, c_{j_1}) = \\
\sum_{i_b} k_{i_b} \sum_{i_b} k_{i_b} &= [\sum_{i_b} k_{i_b}]^2 \\
\sum_{ij} k_{i/j_1} k_{j/i_2} \delta(c_i, c_j) &= \sum_i k_{i/j_1} \sum_j k_{j/i_2} \delta(c_i, c_j) = \sum_{i_2} k_{i_2/j_1} \sum_{j_1} k_{j_1/i_2} \delta(c_{i_2}, c_{j_1}) = \\
\sum_{j_b} k_{j_b} \sum_{i_b} k_{i_b} & \\
\sum_{ij} k_{i/j_2} k_{j/i_1} \delta(c_i, c_j) &= \sum_i k_{i/j_2} \sum_j k_{j/i_1} \delta(c_i, c_j) = \sum_{i_1} k_{i_1/j_2} \sum_{j_2} k_{j_2/i_1} \delta(c_{i_1}, c_{j_2}) = \\
\sum_{i_b} k_{i_b} \sum_{j_b} k_{j_b} &
\end{aligned}$$

where $j_b = i_2 - r = j_2 - r$, $i_b = i_1 = j_1$, $u_{i_b} \in c$ and $v_{i_b} \in c$ these last two conditions can also be formalized with $\delta(c_{i_b}, c_{j_b}) = 1$ if u_{i_b} and v_{i_b} belong to the same community c and $\delta(c_{i_b}, c_{j_b}) = 0$ otherwise.

This development yields :

$$\sum_{ij} k_i k_j = [\sum_{j_b} k_{j_b}]^2 + [\sum_{i_b} k_{i_b}]^2 + 2[\sum_{j_b} k_{j_b}][\sum_{i_b} k_{i_b}] = \sum_{i_b j_b} (k_{i_b} + k_{j_b})^2 \text{ and:}$$

$$\sum_{ij} k_i k_j \delta(c_i, c_j) = \sum_{i_b j_b} (k_{i_b} + k_{j_b})^2 \delta(c_{i_b}, c_{j_b}) \quad (20)$$

Equation (20) can be rewritten using the degrees of nodes:

$\sum_{i_b} k_{i_b}$ is the sum of the degrees of nodes u_{i_b} belonging to c under the condition δ in equation (20). We denote this $d_{u|c}$.

$\sum_{j_b} k_{j_b}$ is the sum of the degrees of nodes v_{j_b} belonging to c under the condition δ in equation (20) and has been called $d_{v|c}$.

$$\text{Then } \sum_{ij} k_i k_j \delta(c_i, c_j) = (d_{u|c} + d_{v|c})^2 \quad (21)$$

C. Analysing first part in (13)

First part in Q is $\sum_{ij} A'_{ij}$. Let's examine what it represents in terms of B . It is possible to identify matrix B in A using indices i_1 and j_2 . Conversely B^t can be identified with indices i_2 and j_1 :

For $i = i_1$ A_{ij} s only produce values for $j = j_2$, moreover for $i = i_2$ A'_{ij} s only produce values for $j = j_1$ with $A'_{i_1 j_2} = B_{i_b j_b}$ and $A'_{i_2 j_1} = B_{i_b j_b}^t$ under typical conditions regarding indices.

$$\text{Then } \sum_{ij} A'_{ij} = \sum_{i_1 j_2} A'_{i_1 j_2} + \sum_{i_2 j_1} A'_{i_2 j_1}$$

$$\text{And } \sum_{ij} A'_{ij} \delta(c_i, c_j) = \sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) + \sum_{i_2 j_1} A'_{i_2 j_1} \delta(c_{i_2}, c_{j_1})$$

The left-hand side of the sum equals the number of edges from nodes u to nodes v inside c .

The right-hand side is the number of edges from these same nodes v and u inside c .

This set-up then leads to:

$$\sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) = \sum_{i_2 j_1} A'_{i_2 j_1} \delta(c_{i_2}, c_{j_1}) \text{ with } i_1 = j_2 \text{ and } i_2 = j_1$$

$$\text{Then } \sum_{ij} A'_{ij} \delta(c_i, c_j) = 2 \sum_{i_1 j_2} A'_{i_1 j_2} \delta(c_{i_1}, c_{j_2}) = 2 \sum_{i_b j_b} B_{i_b j_b} \delta(c_{i_b}, c_{j_b}) \quad (22)$$

This value can also be formalized using the number of edges:

$$\sum_{i_b j_b} B_{i_b j_b} \delta(c_{i_b}, c_{j_b}) = |(u_{i_b|c}, v_{j_b|c})| = |e_{i_b|c, j_b|c}| \text{ where } e_{i_b|c, j_b|c} \in E \text{ \& } u_{i_b|c}, v_{j_b|c} \in c \quad (23)$$

For the entire matrix $A' : \sum_{ij} A'_{ij} = 2 \sum_{i_b j_b} B_{i_b j_b}$

From equation (1), $m = 1/2 \sum_{ij} A'_{ij}$

Let's now define $m_b = \sum_{i_b j_b} B_{i_b j_b} = |e_{i_b j_b}|$ where $e_{i_b j_b} \in E$

Then $m = \frac{1}{2} \times \sum_{ij} A'_{ij} = \frac{1}{2} \times 2 \times \sum_{i_b j_b} B_{i_b j_b} = m_b$

D. Bimodularity

Lastly, by removing sub-index b , which had only been introduced to distinguish indices i and j when applied to A' or B , we can redefine the A' modularity in terms of B :

$$Q = \frac{1}{m} \sum_{ij} [B_{ij} - \frac{(k_i + k_j)^2}{4m}] \delta(c_i, c_j) \quad (24)$$

In terms of edges, by simplifying $e_{i_b|c, j_b|c}$ as e_c (where e_c has both ends in c) and by dropping sub-index b Equation (24) becomes:

$$Q = \sum_c [\frac{|e_c|}{m} - (\frac{(d_{u|c} + d_{v|c})}{2 \times m})^2] \quad (25)$$

The term **bimodularity** refers to this definition of modularity for bipartite graphs since both types of nodes are bound. In previous sections, we have validated the above results on the basis of another author's graph modularity models. It can thus be concluded that (24) offers a good candidate for bipartite graph modularity that takes some specific characteristics into account.